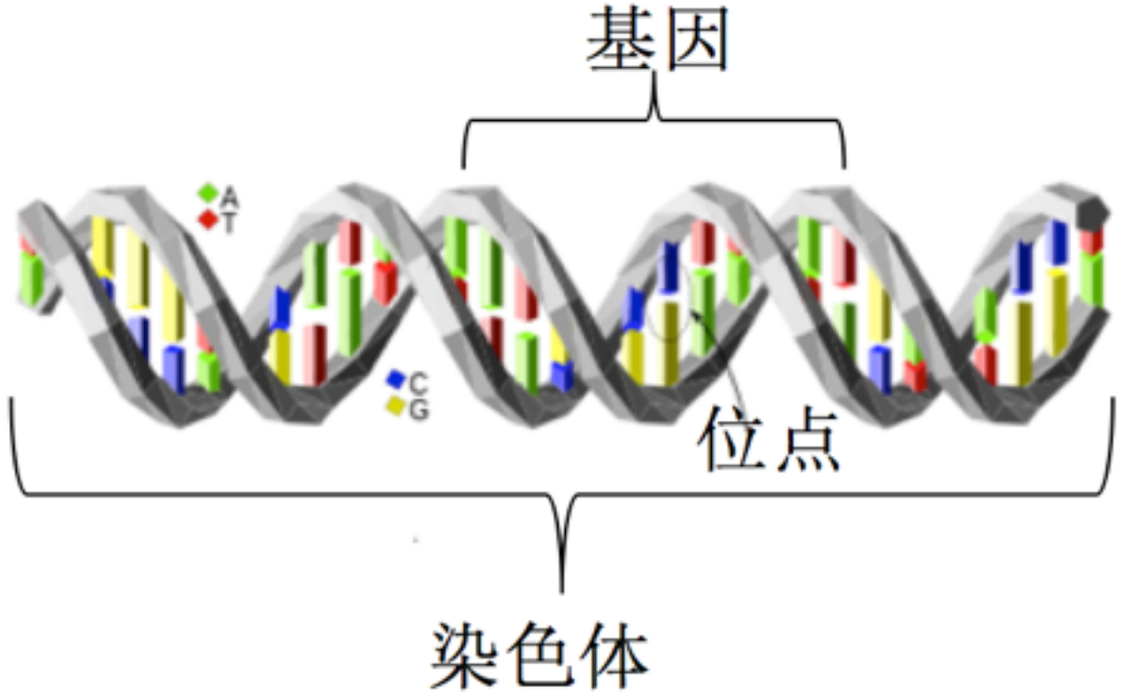
# 具有遗传性疾病和性状的遗传位点和基因分析

强思维　王海洋　李龙元

# 问题背景

- 目标：定位与性状或疾病相关联的位点在染色体或基因中的位置




位点：一些特定位置的单个核苷酸经常发生变异引起DNA的多态性

# 样本数据

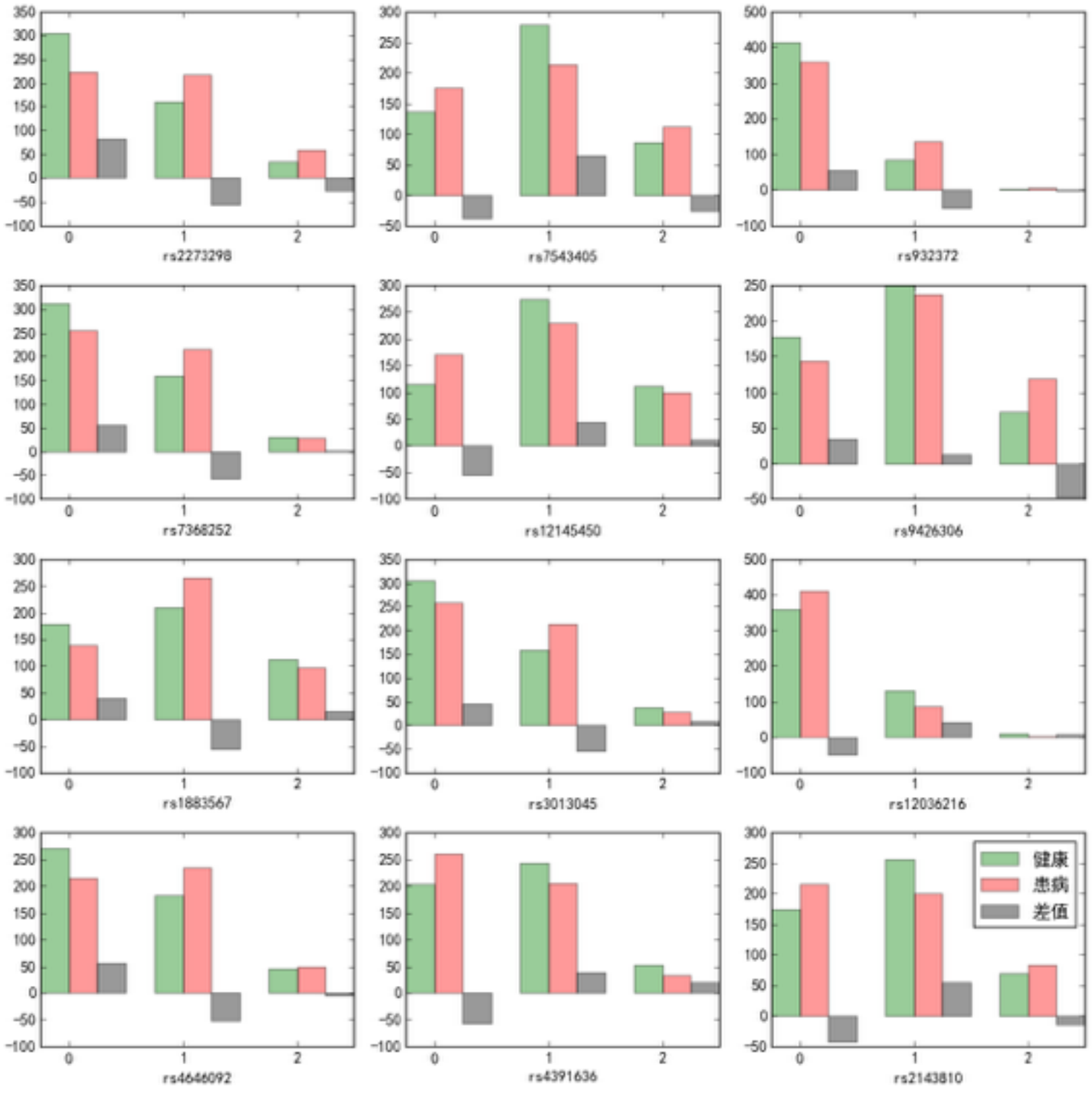
| 样本编号 | 样本健康状况 | 染色体片段位点名称和位点等位基因信息 | | | |
|---|---|---|---|---|---|
| | | rs100015 | rs56341 | ... | rs21132 |
| 1 | 1 | TT | CA | ... | GT |
| 2 | 0 | TT | CC | ... | GG |
| 3 | 1 | TC | CC | ... | GG |
| 4 | 1 | TC | CA | ... | GG |
| 5 | 0 | CC | CC | ... | GG |
| 6 | 0 | TT | CC | ... | GG |

# 问题一

## 碱基(A,T,C,G)编码方式

- 问题分析
- 数值编码
- 类别编码

# 编码方式




健康与患病

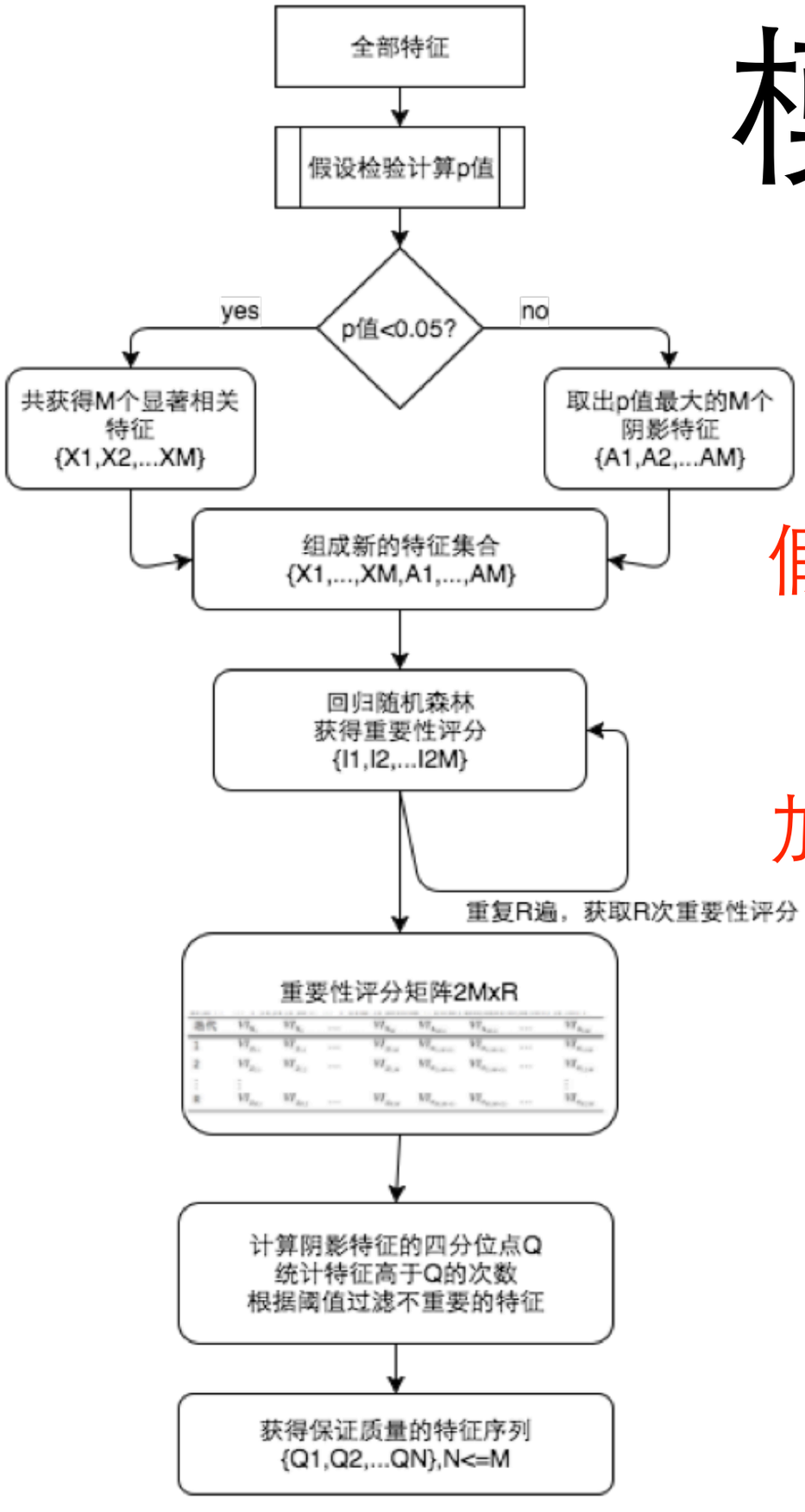
碱基对分布

# 问题二

## 找出某种疾病最相关的一个或几个致病位点

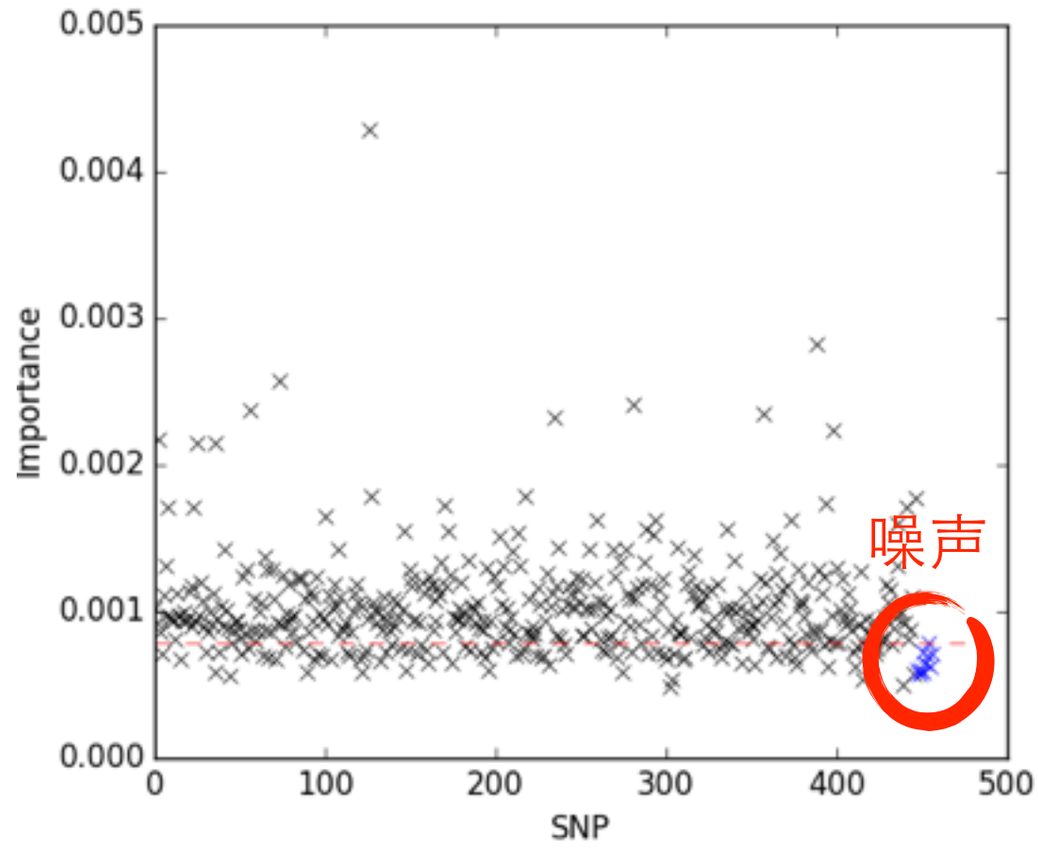
- 问题分析
  - 特征选择方法（子集搜索、过滤、包裹、嵌入）
- 所采用方法
  - 具有质量保证的特征选择方法

# 模型建立



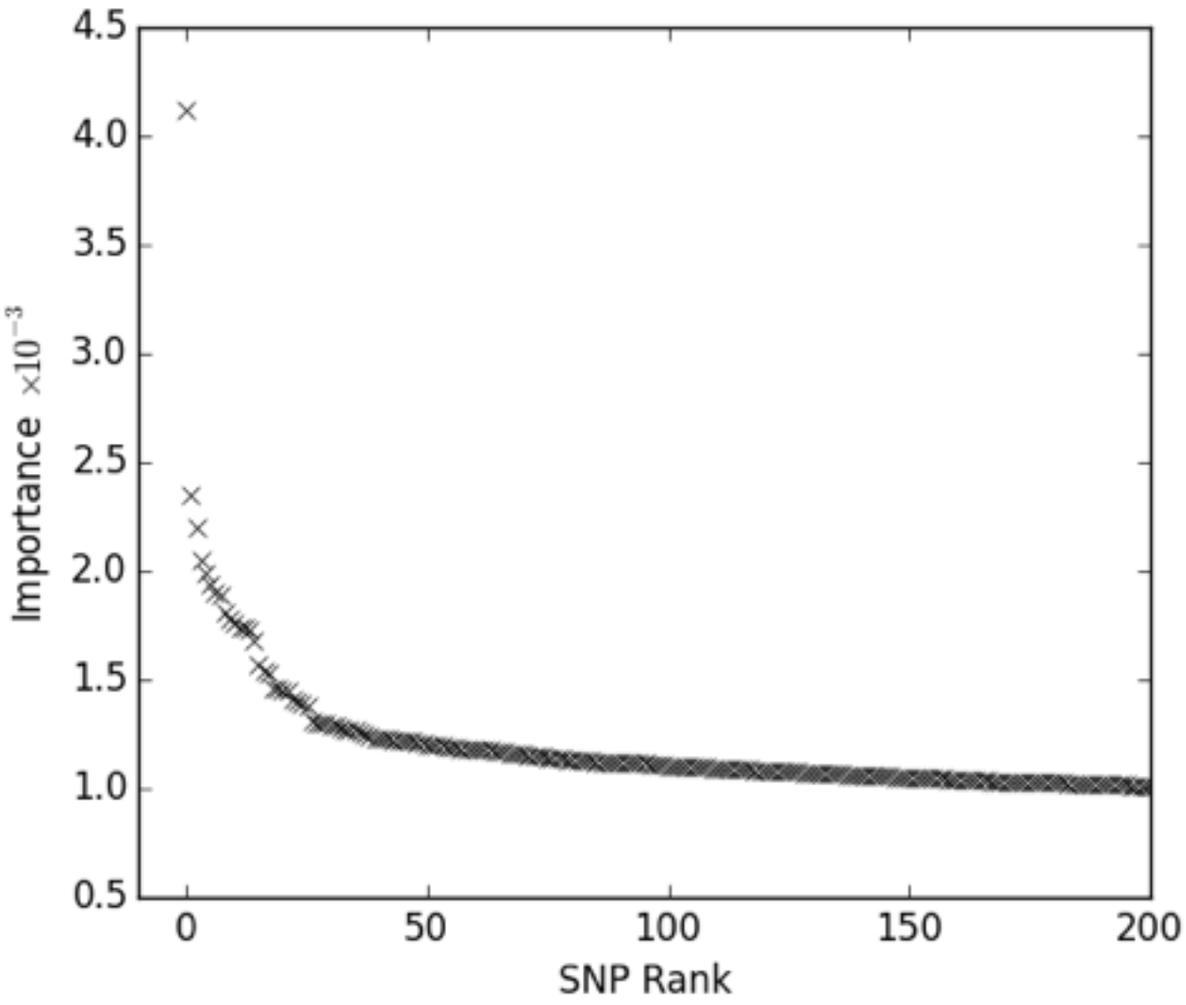

假设检验

加噪评估

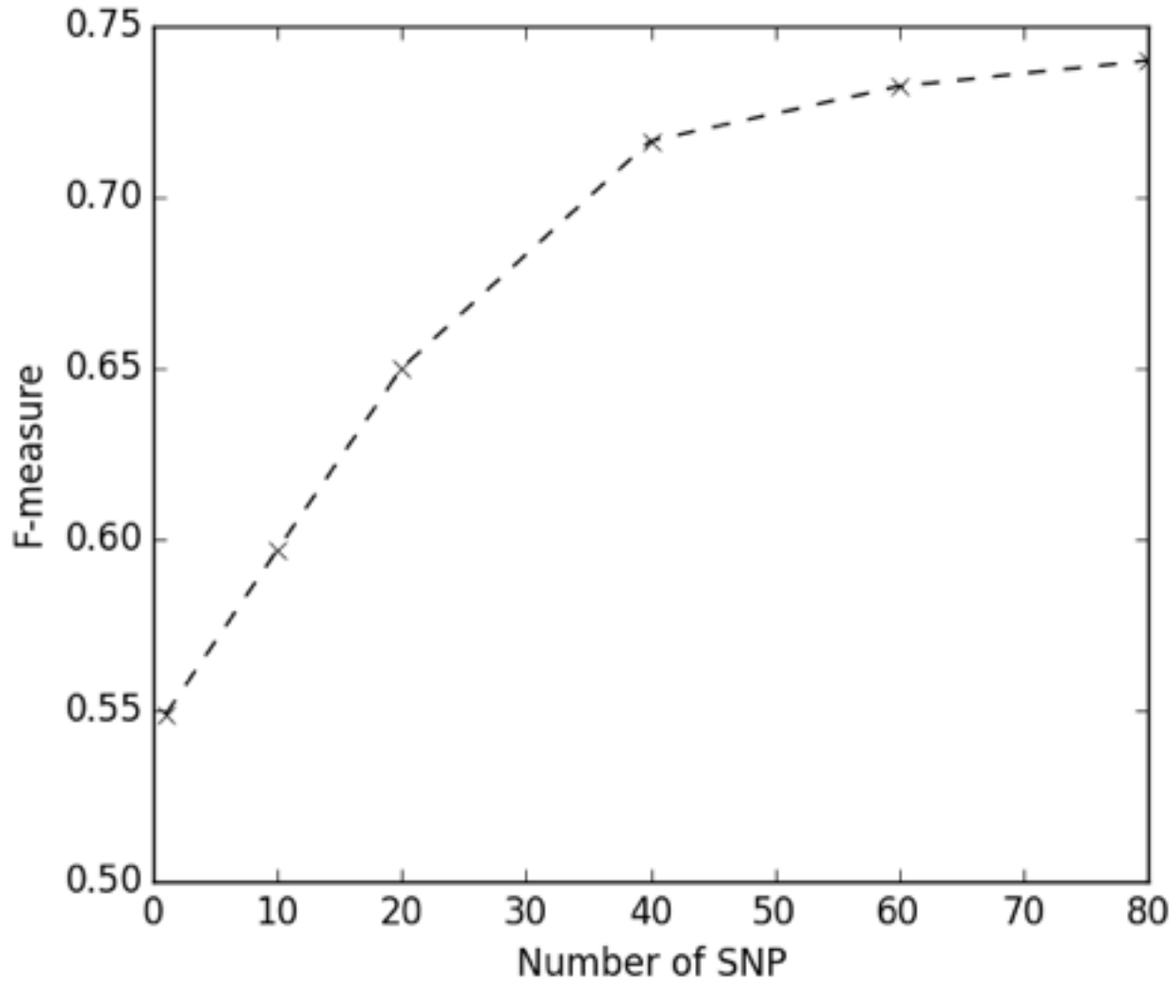



加噪过程

# 模型结果




所选特征




特征分类效果评估

# 模型评估

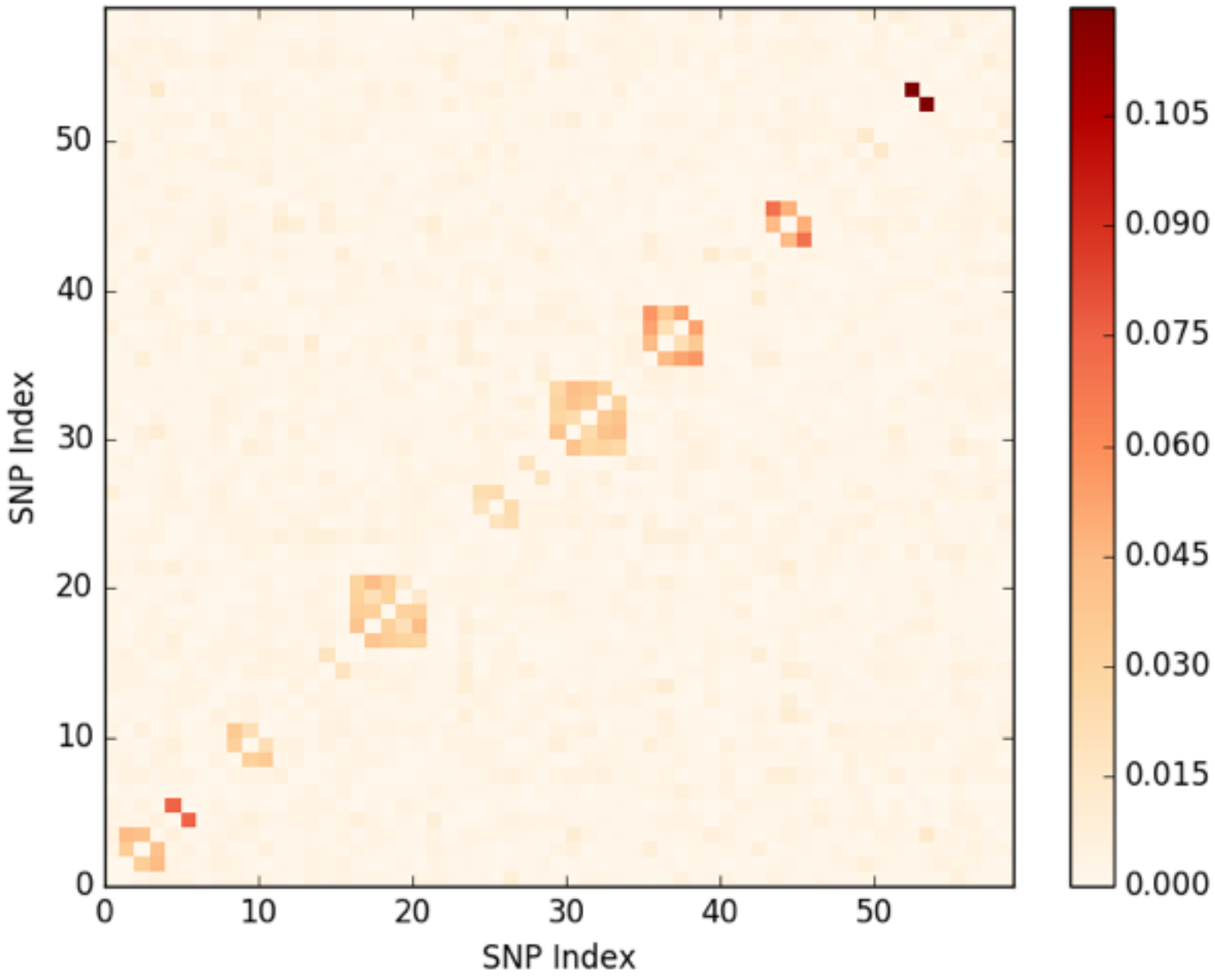
| 模型 | | 准确度 | 召回率 | F1 度量 |
|---|---|---|---|---|
| SVM | SVM(rbf) | 0.656 | 0.6425 | 0.6492 |
| | SVM(linear) | 0.668 | 0.6506 | 0.6592 |
| Tree-based | DecisionTree | 0.552 | 0.5692 | 0.5605 |
| | ExtraTrees | 0.704 | 0.6937 | 0.6988 |
| | RandomForest | 0.712 | 0.7208 | 0.7164 |
| Linear | SGDClassifier(L1) | 0.711 | 0.6985 | 0.7046 |
| | SGDClassifier(L2) | 0.714 | 0.6728 | 0.6928 |
| | LogisticRegression(L1) | 0.711 | 0.7129 | 0.7119 |
| | LogisticRegression(L2) | 0.706 | 0.7034 | 0.7047 |
| Bayes | BernoulliNB | 0.721 | 0.7055 | 0.7137 |
| Ensemble | AdaBoost | 0.722 | 0.7013 | 0.7114 |
| | GradientBoosting | 0.692 | 0.6945 | 0.6932 |

# 问题三

## 找出某种疾病最相关的一个或几个致病基因

- 问题分析
  - 基因由位点的全集或子集构成
  - 常用方法：最(次)显著SNP法、组合法、回归法
    主成分分析、傅里叶分析
- 所采用方法
  - 回归分析法（逻辑回归、岭回归）

# 连锁不平衡效应




互信息度量

同基因不同SNP

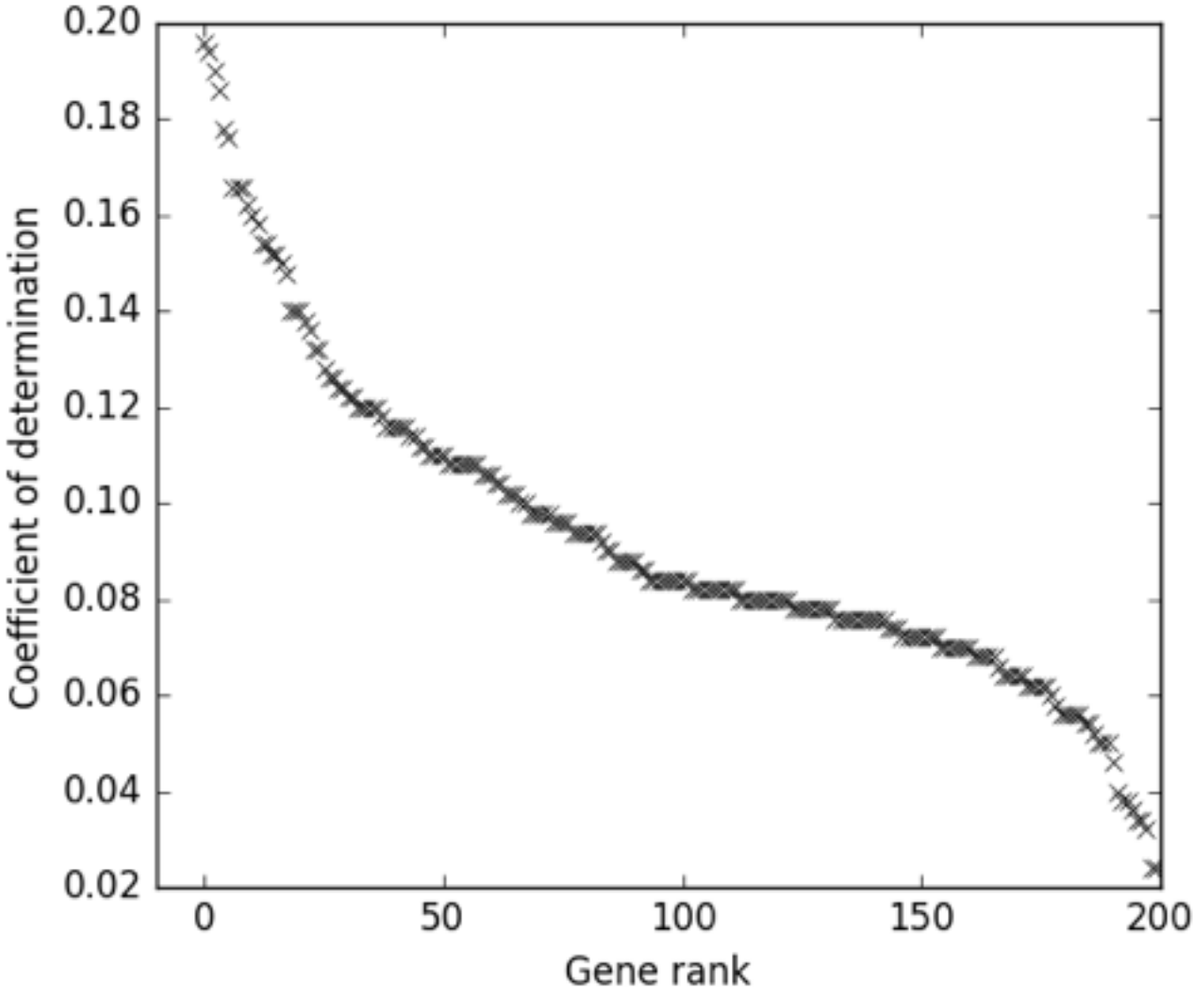
之间存在互信息

# 模型建立

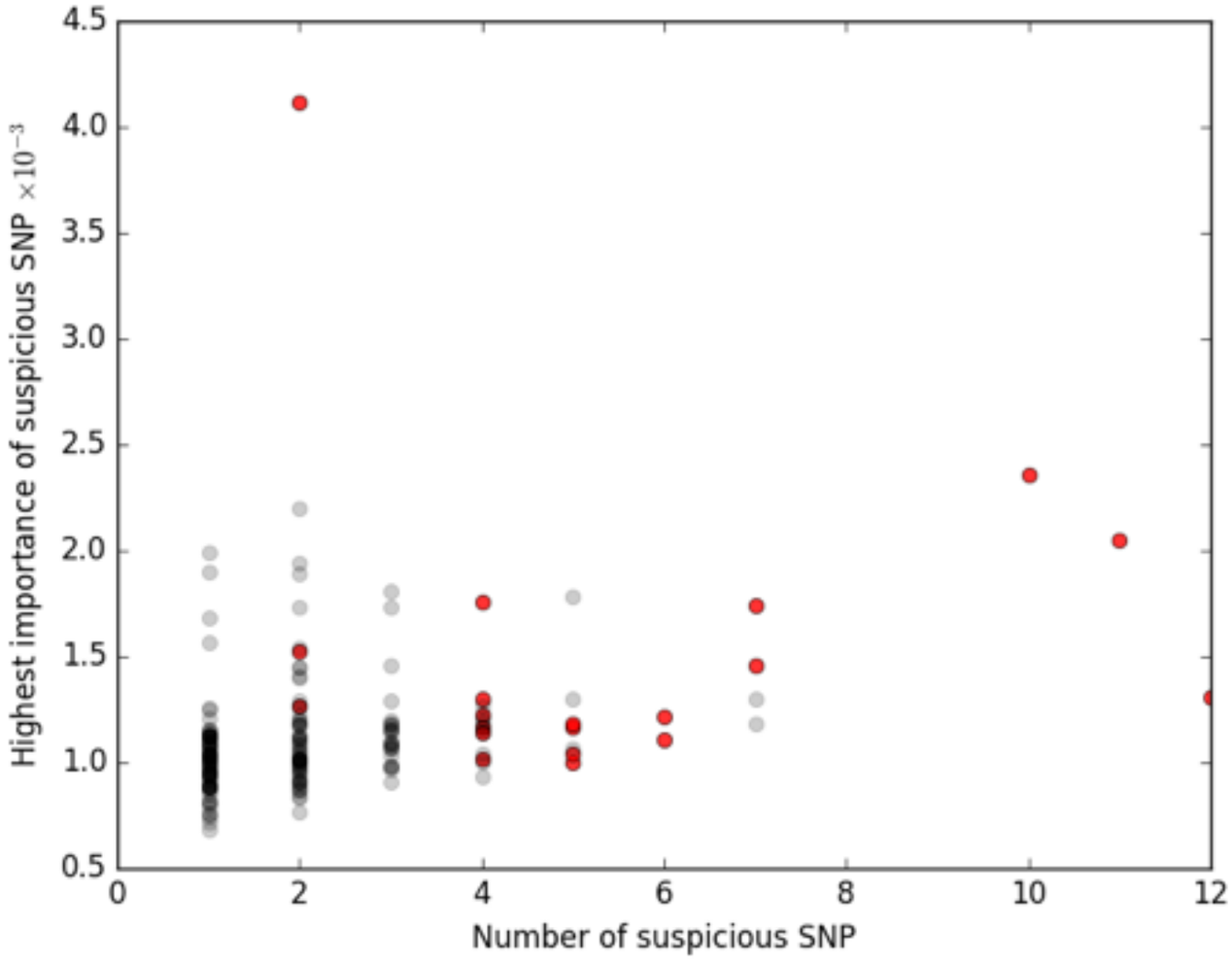
- 逻辑回归
- 岭回归（处理多重共线性）
- 方差分析（决定系数 R-square）

$$R^2=\frac{SSReg}{SST}=1-\frac{SSE}{SST}$$

# 模型结果




所选特征




结果解释

# 模型评估

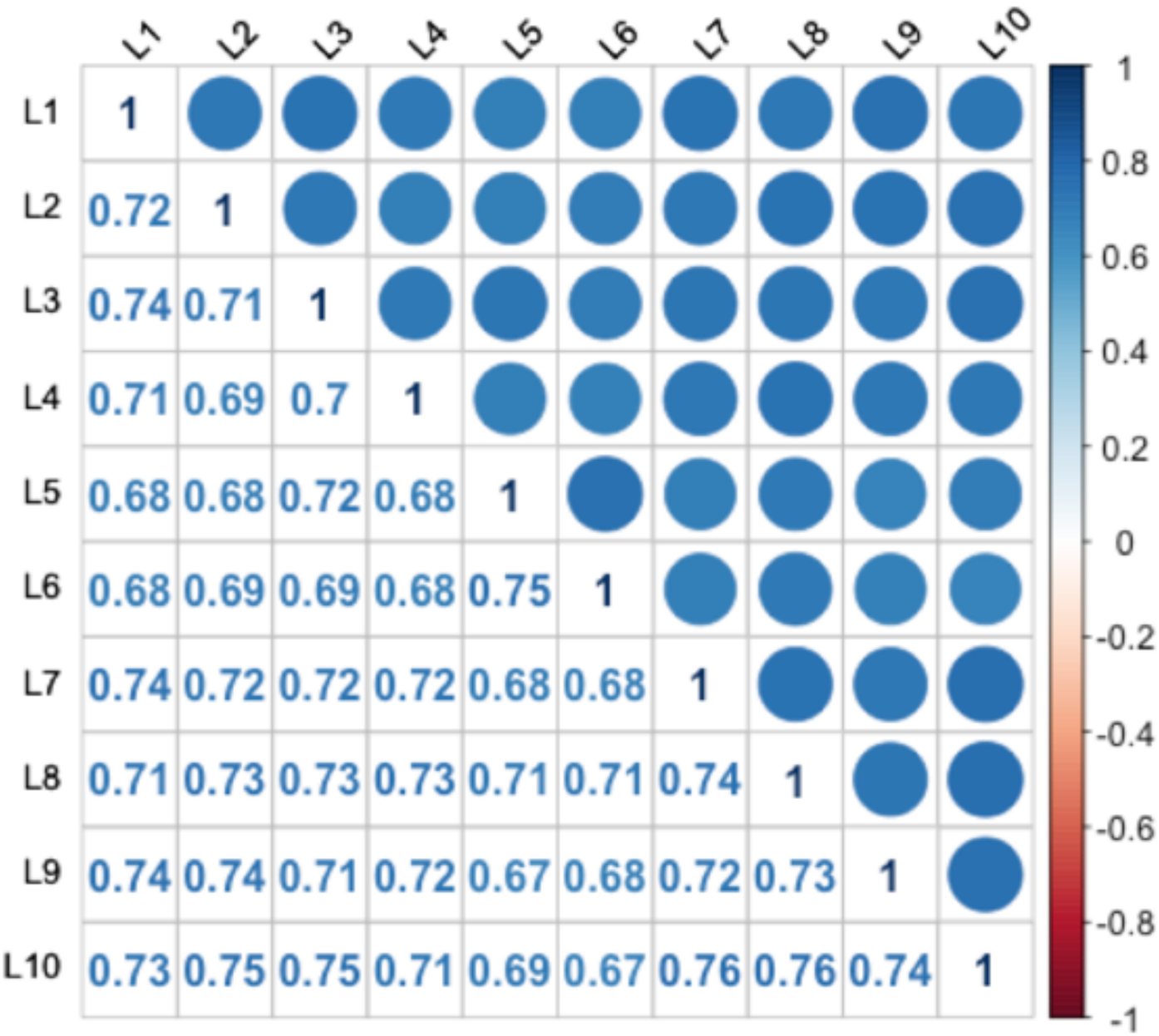
| | 模型 | 准确度 | 召回率 | F1 统计量 |
|---|---|---|---|---|
| SVM | SVM(rbf) | 0.636 | 0.6668 | 0.651 |
| | SVM(linear) | 0.626 | 0.676 | 0.65 |
| Tree-based | DecisionTree | 0.62 | 0.5949 | 0.6072 |
| | ExtraTrees | 0.691 | 0.6873 | 0.6891 |
| | RandomForest | 0.702 | 0.7108 | 0.7064 |
| Linear | SGDClassifier(L1) | 0.679 | 0.6916 | 0.6852 |
| | SGDClassifier(L2) | 0.649 | 0.6827 | 0.6654 |
| | LogisticRegression(L1) | 0.691 | 0.7092 | 0.6999 |
| | LogisticRegression(L2) | 0.654 | 0.6834 | 0.6683 |
| Bayes | BernoulliNB | 0.689 | 0.6951 | 0.692 |
| Ensemble | AdaBoost | 0.684 | 0.684 | 0.684 |
| | GradientBoosting | 0.656 | 0.681 | 0.6682 |

# 问题四

## 找出与多性状最相关的一个或几个致病位点

- 问题分析
  - 性状之间相关性
  - 多标签分类问题
- 所采用方法
  - 多标签随机森林

# 多性状之间相关性




皮尔逊相关系数

均为正相关

相关性趋近

# 模型建立

（1） 给定阈值$\lambda$，将重要的特征和不重要的特征分类两组，$\mathbf{X}_{low}$和$\mathbf{X}_{high}$。

（2） 有放回地抽样训练集$\mathbb{L}$，生成 K 份 Bagged 样本$\mathbb{L}_1,\mathbb{L}_2,...,\mathbb{L}_K$。

（3） 对于每个样本$\mathbb{L}_k$，通过以下方法生成回归树$T_k$

a) 在每个节点，随机选择$m=\left\lfloor\sqrt{M}\right\rfloor$个特征，然后区别为$X_{low}$和$X_{high}$，然后用子空间特征作为划分节点的候选人。

b) 每个数以不确定的方式生成，在每个叶节点，保持叶节点所有$Y$值。

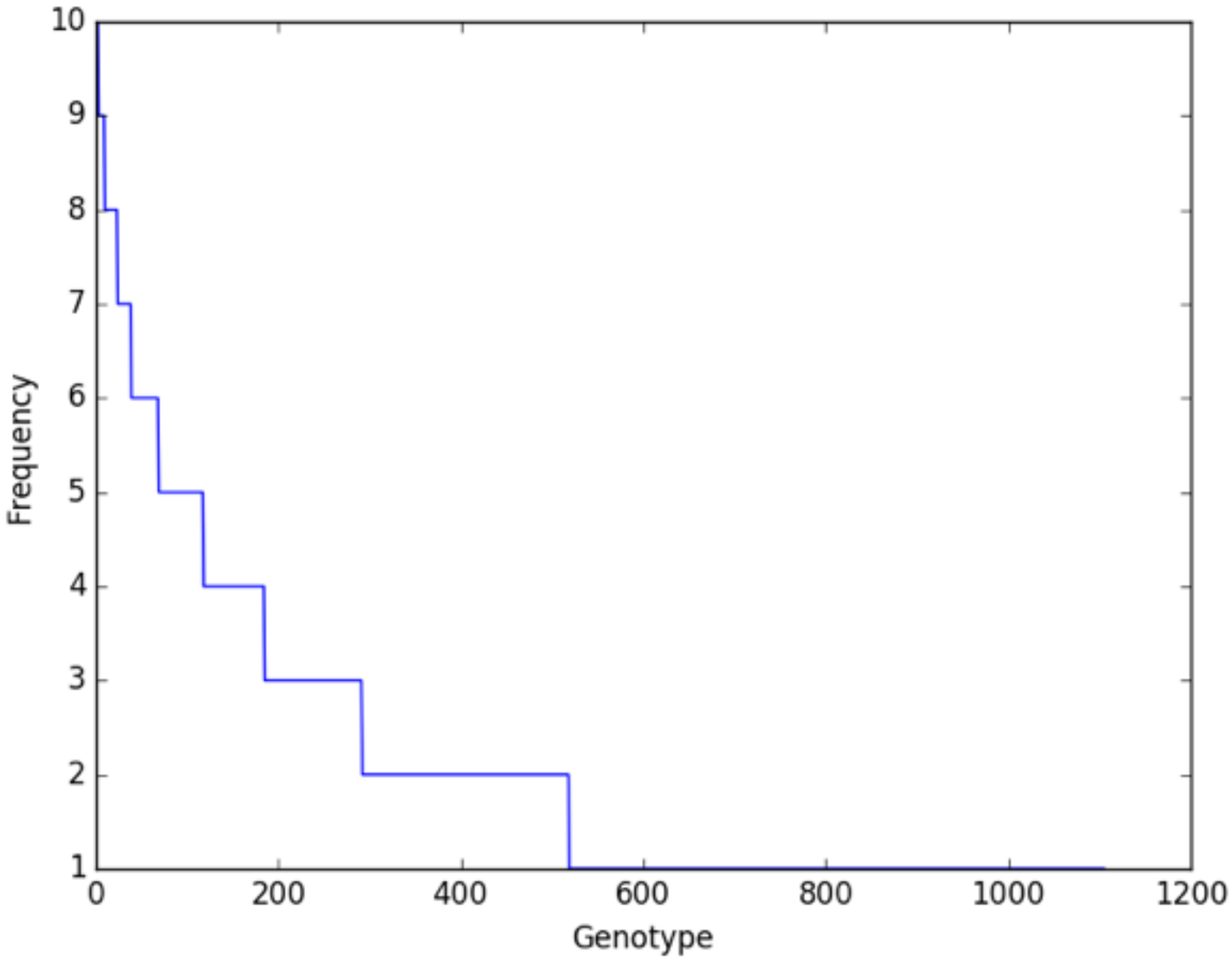
c) 为每棵树和森林的 out-of-bag 样本，计算每个$X_i$的权重。

（4） 给定一个概率$\tau$,$\alpha_l$和$\alpha_h$，使得$\alpha_h-\alpha_l=\tau$，计算相应的四分位点$Q_{\alpha l}$和$Q_{\alpha h}$。

（5） 统计每个特征$\mathbf{X}_i$重要性评分小于 Q 的次数，生成频率列表，取频率高于阈值$\lambda$的特征为最重要的特征。

$$Entropy(D)=-\sum_{j=1}^{q}\left(p(\lambda_j)\log p(\lambda_j)+q(\lambda_j)\log p(\lambda_j)\right)$$

# 模型结果




多种性状

被选频次

# 模型评估

| 建模\结果 | 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 | 9 | 10 | 平均 |
|---|---|---|---|---|---|---|---|---|---|---|---|
| 性状 1 | **0.759** | 0.639 | 0.65 | 0.672 | 0.643 | 0.65 | 0.666 | 0.677 | 0.66 | 0.67 | 0.668 |
| 性状 2 | 0.642 | **0.747** | 0.638 | 0.624 | 0.635 | 0.649 | 0.633 | 0.649 | 0.672 | 0.671 | 0.656 |
| 性状 3 | 0.645 | 0.659 | **0.743** | 0.646 | 0.644 | 0.651 | 0.673 | 0.649 | 0.66 | 0.671 | 0.664 |
| 性状 4 | 0.665 | 0.66 | 0.655 | **0.77** | 0.65 | 0.666 | 0.67 | 0.688 | 0.666 | 0.677 | 0.6767 |
| 性状 5 | 0.654 | 0.634 | 0.64 | 0.646 | **0.742** | 0.664 | 0.642 | 0.634 | 0.653 | 0.637 | 0.655 |
| 性状 6 | 0.656 | 0.654 | 0.654 | 0.655 | 0.678 | **0.749** | 0.653 | 0.648 | 0.662 | 0.66 | 0.667 |
| 性状 7 | 0.67 | 0.654 | 0.669 | 0.685 | 0.643 | 0.652 | **0.76** | 0.686 | 0.688 | 0.689 | 0.68 |
| 性状 8 | 0.686 | 0.656 | 0.673 | 0.68 | 0.654 | 0.649 | 0.686 | **0.749** | 0.671 | 0.662 | 0.677 |
| 性状 9 | 0.667 | 0.677 | 0.674 | 0.66 | 0.651 | 0.631 | 0.676 | 0.67 | **0.757** | 0.666 | 0.673 |
| 性状 10 | 0.66 | 0.666 | 0.664 | 0.656 | 0.649 | 0.643 | 0.66 | 0.67 | 0.66 | **0.753** | 0.668 |
| **多标签** | 0.706 | 0.721 | 0.722 | 0.743 | 0.726 | 0.707 | 0.719 | 0.713 | 0.737 | 0.72 | **0.721** |

Thank you!